High Performance Computing Facilities for the Next Millennium

Dealing with New Technology

SC99 Tutorial
November 14, 1999

Tammy Welcome

Advanced Systems, Group Leader

tswelcome@lbl.gov

Best Value Source Selection (BVSS) Provides Flexibility to Get Best Solution

- High level (no detailed SOW in RFP)
- Baseline requirements
  - Establish minimum requirements to be considered responsive
- Value-related characteristics
  - Qualitative criteria for subjective evaluation of proposals
  - We provided list in RFP and RFP asked Vendor to identify others
- Result - let Vendor design their system

Four Criteria Were Used to Determine System That Best Met NERSC Requirements

- **Feasibility**
  - **Likelihood of success, balanced plan, manageable solution**
- **Applicability**
  - **Increase in computational capability, production system, satisfies NERSC goals**
- **Capability**
  - **Corporate commitment, state-of-the-art, how will management and personnel ensure success**
- **Affordability**
  - **Cost effective, meets NERSC budget constraints**

Timeline of Events

- Technology survey
- Originating requirements
- Validate requirements/feasibility
- Pre-release of benchmark
- Release RFP/test suite
- Responses received
- Evaluation
- Negotiation
- Initial Delivery
- Entire Process Took 1.5 Years From Technology Survey and Requirements Gathering Until Delivery of System
- Factor in reviews, approvals, financing, holidays!

First Things First

- Select right people for team
- Identify member roles and responsibilities
- Understand process
- Have goal in mind
- Keep running cache of overheads

Benchmark Preparation Requires Significant Time and Manpower

- Select "star" benchmark that represent the future not the past
- Limit size and complexity of benchmark suite
- Have strict guidelines for benchmark selection and preparation
- Be clear and explicit about benchmark instructions
- Pre-release benchmarks 3 months prior to RFP
- Use web
- Focus on what's important

Carefully Prepare RFP

- Hold offsite meetings to create RFP
- Make it clear to reader what is important and what is not
- Determine all phasing strategies and options at this time
- Include facilities requirements
- Ensure no requirement "surprises" from other parts of organization
- Describe clearly negotiation expectations (detailed SOW)
- Provide spreadsheets for system configuration, benchmark results
- Use Web
- Consider pre-solicitation conference or pre-release of draft documents to clarify RFP and benchmark instructions prior to releasing final version
- FOCUS

Focus on What is Important in Evaluating Proposals

- Hold offsite evaluation meetings
- Triage the data provided and proposals
- Use additional outside information (contacts and other sites, papers, conferences, etc.) to aid in the evaluation
- Make use of spreadsheets
- FOCUS

Preparation of SOW During Negotiations Helps with Disclosure of Information

- Hold offsite negotiation meetings
- Insist that everyone is present who has stake in negotiation outcome
- Present draft SOW before negotiations start
- Have one person maintain revision control of SOW
- Make use of spreadsheets
- FOCUS

Recommend BVSS for Use in Procurement of Large Systems

- Gives vendor flexibility to be creative in meeting your requirements
- Gives organization flexibility to choose system that provides best value

Detailed Evaluation of Baseline Requirements and Value-Related

- **Performance (sustained performance, network and file system I/O, individual benchmarks results)**
- **User environment (programming environment, enhance and integrate with existing environment, roadmap, documentation, training, standards, life cycle cost, functionality and ease of use)**
- **System management (checkpoint/restart, OS related software, design and implementation of integrated system, roadmap, standards, life cycle cost, functionality and ease of use)**
- **Reliability (repair response plan, MTTR/MTBI/MTBF, reliability of service, maintenance**
- **Corporate commitment (milestone schedule, key people, management and corporate capability, ability to meet schedule, ability to test and produce system, options offered)**
- **Facilities (power, space, schedule/delivery)**

Linux Clusters

- Why are we talking about Linux clusters?
  - How do they compare to NERSC MPP?
  - NERSC looking to roadmap for future.
- What is NERSC currently doing?
  - Production Cluster
  - Research Clusters
  - Software R & D
- What questions are before us now?
  - Workload —Resources
  - Tradeoffs —Technical
- Should NERSC be running large production LINUX cluster for general user community?

Why are we talking about Linux Clusters?

- **Linux stability and acceptance has made real strides in the last 18 months. (eg. User Survey)**
- **NERSC has shown that a production Linux cluster is feasible.**
- **NERSC has software/hardware projects w/ direct applicability to development of Linux clusters.**
- **ASCI's "above the line" vs "below the line"**
- **Question: Do Linux clusters offer best value for a certain class of NERSC workload?**

- ***Time is right to consider where Linux clusters today may lead in the near future and how they can solve computational needs of NERSC users.***

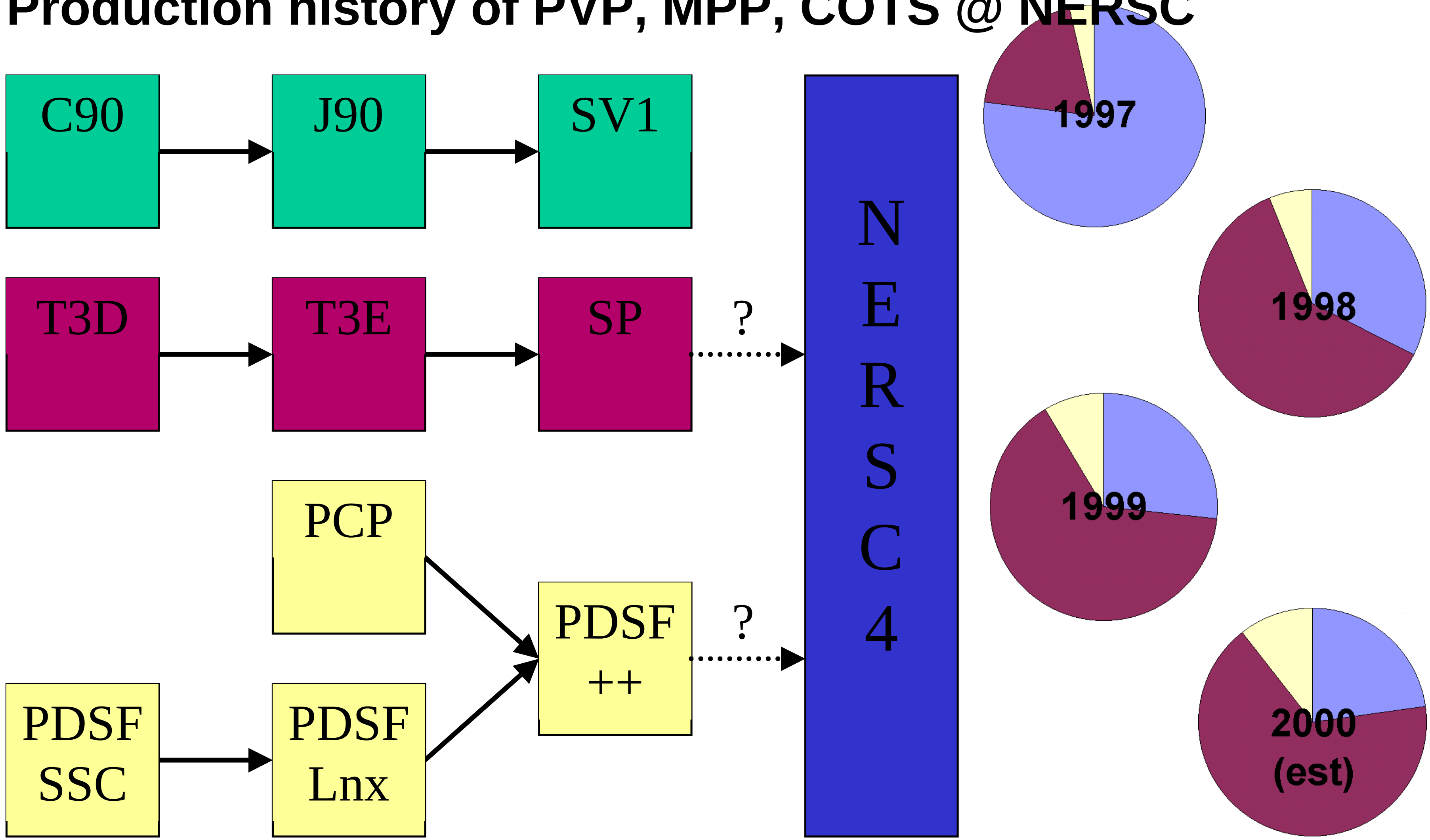
MPP & COTS Linux

- Traditional distinctions are blurring. Still useful to consider.
- Hardware:
  - MPP: Homogeneous nodes
  - COTS: Heterogeneous slices of homogeneous nodes
- System:
  - MPP: Single System Image
  - COTS: Multiple identical systems
- Network Interconnect:
  - MPP: Fast, proprietary
  - COTS: Slow, commercial

MPP & COTS Linux

- File System:
  - MPP: Global
  - COTS: Shared + Local
- N-Way Jobs:
  - MPP: N-way job requires N CPUs
  - COTS: 1 node down does not stop N-way job
    - (FARM-like Workload)
- Space, Cooling, Power Requirements:
  - MPP: Densely Packed - Less space, more power, more cooling
  - COTS: Loosely Packed - More space, less power, less cooling

Production history of PVP, MPP, COTS @ NERSC



What is NERSC currently doing?

- Linux Clusters
  - Production - PDSF
  - Research - PCP, Babel
- File Systems & Storage
  - NFS
  - HPSS
  - DPSS
  - GFS
- Communications SW
  - VIA: M-VIA
  - MPI: MVICH
- Inter-Institution Projects
  - High-End Cluster SW
  - Scalable GPFS
- Production Environment
  - BLD - Berkeley Lab Dist
    - Cluster & Farm
  - Batch/Load Sharing
    - LSF, PBS, Mosix
  - Administration & Management Tools
  - Performance Studies
- CPU Hardware
  - Intel, Alpha, Solaris
- Network Hardware
  - 100bT, 1000bT, Myrinet, Etherchannel, Giganet, ServerNet

PDSF - 100 CPU Production Cluster

- PDSF - Parallel Distributed Systems Facility
  - HENP community
    - Specialized needs/Specialized requirements
    - 30 groups, 280 users
- Intel Linux batch & interactive CPUs
  - 13*PII/266, 16*PII/333, 28*PII/400, 42*PIII/450
  - Linux kernel v2.2.12
- Solaris interactive CPUs (5 UltraSparc)
- NFS Linux Data Vaults
  - 4.2 TB global disk (RAID & non-RAID)
- LSF - Load Sharing Facility
- http://pdsf.nersc.gov/

PDSF Hardware "Projections"

- Current:
  - CPU: 2195 SPECint95
  - DISK: 4.2 TB
  - NET: 100 Mbs
- 2 Year Plan (STAR):
  - CPU: >8000 SPECint95
  - DISK: >16 TB
  - NET: 1000 Mbs
- 4 Year Plan (ATLAS):
  - CPU: ~20000 SPECint95
  - DISK: ~50 TB
  - NET: 1000+ Mbs ?

PCP & Scientific Computing

- NERSC PC Cluster Project (PCP) Goal: Make feasible widespread use of PC clusters for scientific computing.
- Develop software infrastructure for assembling scalable plug-and-play clusters from PCs
- Develop critical enabling software components
- Ensure uniform HPC software environment
- Perform, collect, and disseminate analysis of hardware and software
- 32 Intel (400 MHz PII) Linux CPUs
- http://www.nersc.gov/research/ftg/pcp/

Babel

- Research into high performance communication and cluster software
- Multidisciplinary collaborative research spanning cluster software, grid infrastructure, numerical algorithms, applications, and visualization
- 12 Alpha EV6-based Digital DS10 workstations
- See exhibit in NERSC booth

Optimized Linux NFS

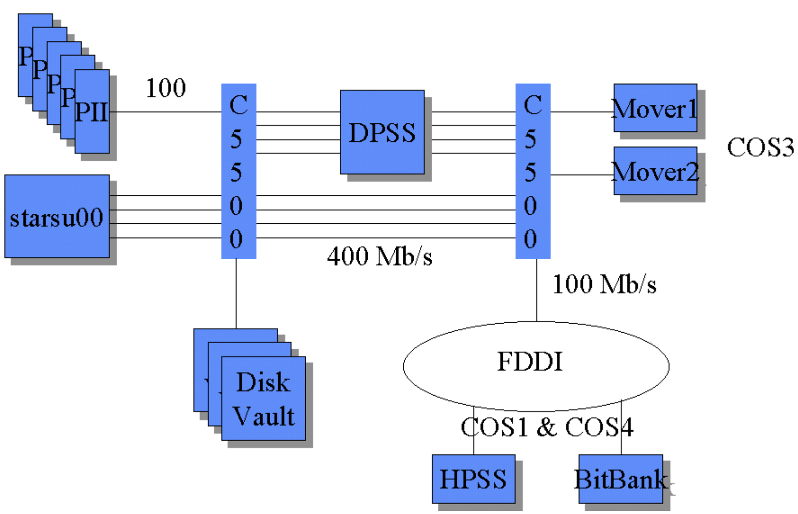
- Built using kernel based NFS servers & large volumes using IDE drives
- Benchmarking of linux v2.2.x NFS client testing of NFS V3 client updating of channel bonding to work w/v2.2 & Cisco
- Combined all together to create Linux based NFS servers capable of sustained 20/mbs network read/write rates.
- Thomas Davis internationally recognized Linux NFS authority.

New PDSF Data Vaults - Performance

- Local (no network) Performance
  - 63 MB/sec read rates (Bonnie block)
  - 13 MB/sec write rates (Bonnie block)
- Network (NFS) Performance
  - one stress test (300 GB) was capable of writing 8 MB/sec (from 7 separate nodes at once) to raid array while the array was building parity.
  - second stress test is writing at rate of about 20MB/sec (rebuild was complete).
  - Network (200 Mbps) is the bottle neck
    - Need to upgrade to Gigabit Ethernet

PDSF Use of HPSS (May'99)

- Data intensive computing:
  - high speed network
  - large, performant, stable mass storage
- > 7 TB of HENP data
  - BNL, CERN, Astro.
- 9.5 MB/s I/O measured



user	files	space	io	SRUs
TOTAL	**6461316**	**81697.7**	**2705.4**	**51260.4**
dbest	37122	2860.5	61.0	1432.8
pdsf	3376	1422.2	29.7	691.6
snelling	49942	1075.1	0.4	491.7
fqwang	64201	803.2	3.6	412.8
zimm	32211	726.8	1.3	334.5
saul	54395	720.1	6.7	380.1
olson	8137	351.6	0.6	152.9
gxrai	29082	251.7	3.5	149.5
liq	1422	137.8	0.0	56.9
partlan	394	131.3	0.0	53.0
odyniec	1808	99.9	0.0	42.1
yangj	3705	99.8	1.3	49.4
dahl	1012	97.9	0.0	40.4
hardtke	81	60.8	0.0	24.4
jacobs	4483	50.0	0.0	25.4
heng	6967	45.6	0.0	26.6
ianh	315	30.8	0.0	12.7
sakrejda	146	9.3	0.0	3.9
nevski	414	3.6	0.0	1.9
may	77	2.7	0.0	1.2
margetis	400	2.2	0.0	1.3
HENP	**299690**	**8982.9**	**108.1**	**4385.1**
HENP(%)	**4.6%**	**11.0%**	**4.0%**	**8.6%**

DPSS Design

- Support specialized data-intensive applications
- Provide very high data throughput
- Parallelism at every level, including disk, SCSI bus, network, and server
- High-speed WAN aware
- Scaleable throughput and capacity
- Economical
  - Use only low-cost commodity hardware components
- Location transparency
  - Location of DPSS servers is transparent to the application

Global File System

- Working on plan to prototype GFS in NERSC environment (proof of concept, hardening, readying for production environment)

- Transfer large amounts of data - Terabytes
- High bandwidth - 500 MB/sec / Terabyte of data
- High availability
- Heterogeneous - (AIX, UNICOS, LINUX, Solaris, FreeBSD...)
- Scalable with multiple streams of data

BLD - Berkeley Lab Distribution

- Software distribution that makes it easier for scientists to turn a collection of PCs into a usable cluster
- Provide key tools for configuring, managing, and running jobs on cluster (task farm and parallel clusters)
- Some early software available, general availability early 2000
- See SC'99 tutorial on production Linux clusters
- http://www.nersc.gov/research/bld

ACTS (Advanced Computational Testing and Simulation) Project

- ACTS toolkit
  - Set of DOE-developed software tools for developing parallel applications
- Toolkit includes:
  - High performance numerical libraries
  - Tools for better code design
  - Tools that enable new classes of technology
- Interoperability of tools is goal of toolkit
- Information and Support Center - Consumer Reports providing descriptions, documentation, evaluations, and advice
- See booth exhibit
- http://acts.nersc.gov

M-VIA for LINUX

- M-VIA is Modular Implementation of Virtual Interface Architecture for LINUX
- VIA features:
  - Provides industry-standard architecture for communication within clusters
- M-VIA features:
  - High performance
  - High portability
  - Robustness
  - Reference implementation
- http://www.nersc.gov/research/ftg/pcp/via/

M-VIA Development

- M-VIA is Research Prototype Undergoing Active Development
- M-VIA 1.0 released September 25, 1999
  - Full robust implementation
  - Small number of drivers
- M-VIA 2.0
  - Improve internal interfaces based on M-VIA 1.0 feedback
  - Developer's release available
  - Large number of drivers (giganet, myrinet, servernet)
- MVICH (MPI over VIA) preliminary version released

What questions are before us now?

- Workload
  - How much/which fraction of NERSC workload is appropriate to consider Linux clusters?
- Resources
  - What magnitude of NERSC resources can be applied to developing production cluster?
- Tradeoffs
  - What would NERSC customer base be willing to give up for production cluster?
- Technical
  - What technical hurtles are still unaddressed?

Question:

- **Before end of 2000, should NERSC be running large production LINUX cluster for general user community?**